Text-mining Tools for Optimizing Community Database Curation Workflows in Neuroscience

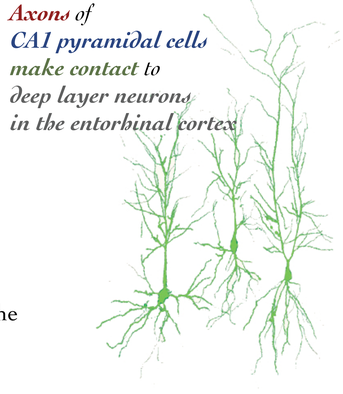
Kyle H. Ambert, B.A.
Oregon Health & Science University
Department of Medical Informatics & Clinical Epidemiology, Portland, Oregon, USA
Supported in part by 1R01LM009501-01 & 5T15LM007088

Abstract

The emphasis of multilevel modeling techniques in the Neurosciences has led to an increased need for large-scale databases containing neuroscientific data. Despite this, such databases are not being populated at a rate commensurate with their demand amongst Computational Neuroscientists. The reasons for this are common to scientific database curation in general--limitation of resources. Much of Neuroscience's long tradition of research is documented in computationally inaccessible formats, such as the pdf, making data extraction laborious and expensive. Here, we propose a series of studies designed to mitigate the bottlenecks in Neuroscience database curation. In particular, we focus our efforts on the Neuron Registry (NR), a community database of neuron-related information pulled from the primary literature. We describe three research projects and how they will extend preliminary research we've already completed to address the needs of the NR. First, we demonstrate how active learning can be used to efficiently increase the volume of data in the NR. Next, we describe the role of document classification algorithms in the NR workflow, discussing the motivation behind the machine learning approaches that were selected. Next, we show how a submission classification system will address important issues inherent to community databases, in particular the NR. Finally, we show how the results of our work here will be relevant to Computational Neuroscience and Biomedical Informatics alike, providing a novel solution to the problem of inefficiency in the development and maintenance of community data resources.

The Neuron Registry (NR) is a community database of neuron types, defined in terms of their properties.

Such a database will be useful for computational studies, which will further our understanding of the neuronal bases of certain diseases.

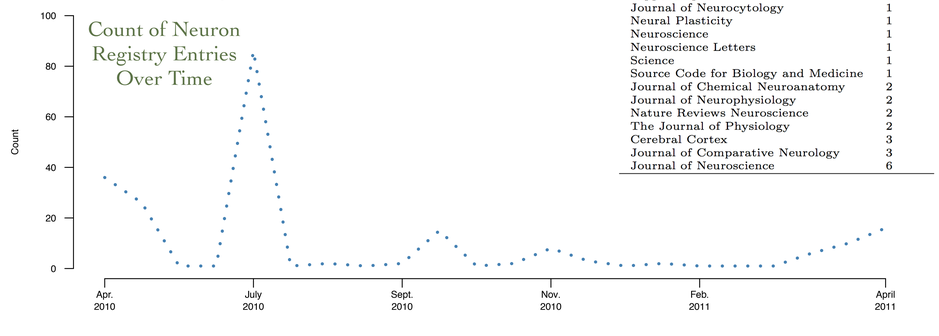


The Challenges:

[1] Use text-mining techniques to efficiently increase the volume of data in the NR.

[2] Identify machine learning techniques useful for classifying the Neuroscience literature.

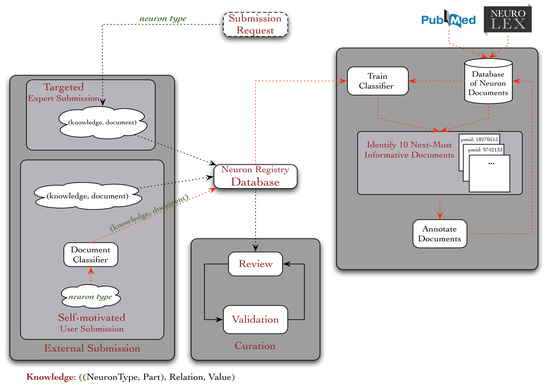
[3] Develop text-mining methods for optimizing the community database curation workflow.



Journal	Count
Brain Research	1
European Journal of Neuroscience	1
Hippocampus	1
Journal of Neurocytology	1
Neural Plasticity	1
Neuroscience	1
Neuroscience Letters	1
Science	1
Source Code for Biology and Medicine	1
Journal of Chemical Neuroanatomy	2
Journal of Neurophysiology	2
Nature Reviews Neuroscience	2
The Journal of Physiology	2
Cerebral Cortex	3
Journal of Comparative Neurology	3
Journal of Neuroscience	6

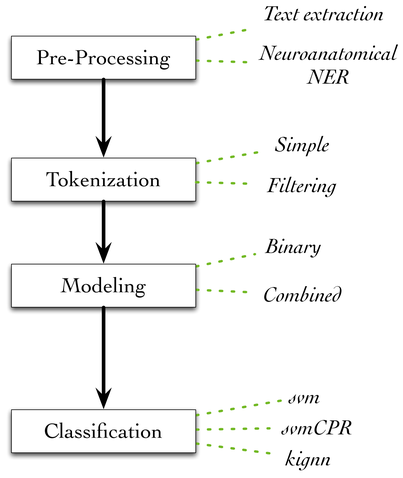
Methods

[1] Increase Data



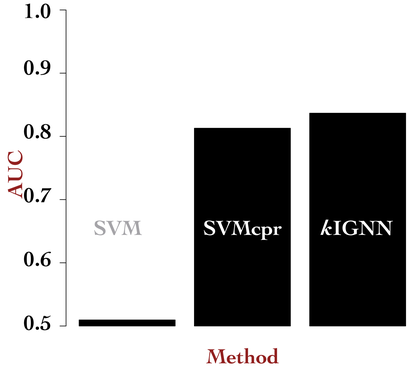
- Active learning will be used to optimize the annotation of additional data.
- Annotation will proceed until:
  - No statistically significant performance change.
  - 50% coverage of neuron types achieved.

[2] Text-mining in the Neuroscience Literature

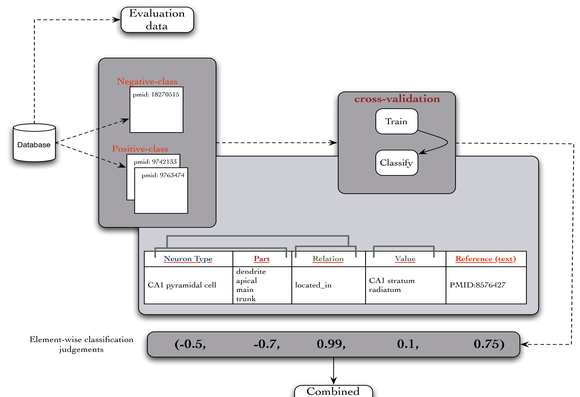


- Preliminary study, using manually-curated data from a related database.
- kIGNN was better able to identify documents for inclusion in the database than competing algorithms.
- A document classification system will be developed for comparing various algorithm combinations.
- 5x2 cross-validation will be used for system development.

AUC by Classification Method



[3] Community Database Curation



- A submission classification system will be developed, to automatically alert curators of unlikely database additions.
- Four types of erroneous submissions will be simulated:
  - Useful document, incorrect knowledge
  - Not useful document, incorrect knowledge
  - Filtered permutations of neuron types, relations, and NR document identifiers
  - Useful document, correct knowledge, non-primary source

Conclusions

[1] We describe a set of studies that will influence the Machine Learning, Biocuration, and Neuroscience communities.

[2] By extending previously-developed text-mining approaches, we will be able to create a text-mining system for community databases that will be useful in both low-data and high-traffic scenarios.